Tim Vigers

3/26/19

Consulting 2

Group Project

*Propensity Scores*

*Introduction*

Randomized controlled trials (RCTs) ensure by design that treatment groups have similar baseline characteristics, but these characteristics can differ significantly in observational studies (for example through participant self-selection). It is important to account for these differences before making inferences on treatment effect based on observational data, since one cannot compare outcomes directly as with a RCT. One common method for dealing with this issue is regression adjustment. [1] Propensity scores, or the probability that a participant is assigned to a particular treatment group given their observed baseline characteristics, are simply another way to “remove bias due to all observed covariates.” [2]

There are a few reasons that propensity scores are used in place of regression adjustment. First, propensity scores allow for separation between the design and analysis phases of the study because “a matched, stratified, or weighted sample can be constructed

without any reference to the outcome.” [1] Also, research has shown that propensity scores are more effective than regression when the outcome under study is rare [3] and slightly better for estimating hazard and odds ratios. [1]

However, once a researcher has decided to use a propensity score approach rather than regression adjustment, there are several questions regarding its implementation. First, how should one estimate the propensity score (which baseline covariates should be considered)? Next, how should one evaluate this score estimate? Finally, what is the best algorithm for matching treatment groups based on propensity score, and how should one assess match quality? [4]

*Methods*

Score model

The most common method for estimating the propensity score is using a logistic model regressing treatment group on a set of baseline characteristics. [1] Lee et al. and Setoguchi et al. examined logistic regression alternatives such as classification and regression trees (CART), random forests, and other machine learning or data mining techniques. [5, 6] However, many of these potential replacements were evaluated using c-index, which “provides no information as to whether the propensity score model has been correctly specified.” [1]

In theory, matching participants by true propensity score will ensure that baseline covariates are equally distributed between groups (i.e. independent of treatment), so evaluation of the score model should assess to what extent this is achieved. One method is to look at the standardized difference, which converts difference in means to units of pooled standard deviation. It is therefore unaffected by sample size and allows for comparison between variables measured in different units, although there is no consensus on a useful threshold for this measure. [1]

If systematic differences in potential cofounders remain (the groups are unbalanced) after conditioning on propensity score, then the score model requires adjustment. Assuming a logistic model, this correction can be performed using standard model building techniques (addition of interactions, non-linear terms, etc.). [1] Caliendo et al. also recommend the hit or miss method, statistical significance testing, leave-one-out cross validation, or overweighting some variables as solutions to poor group balance, [4] although Austin strongly discourages significance testing as p values are confounded by sample size. [1]

Matching

Assuming one is able to develop a reasonable model for the propensity score, the next step is to determine the appropriate matching algorithm. Caliendo et al. recommend five general approaches, all of which involve some sort of trade-off between bias and variance. The easiest and most common method is nearest neighbor (NN) matching, in which each participant from the control group is matched with the participant in the treatment group with the closest propensity score (or vice versa, depending on the numbers in each). This can be done with or without replacement. Matching with replacement can be useful if, for example, there are many more treatment participants with a high propensity score than control participants. In this case each control participant might be matched with multiple treatment individuals. [4] Although this results in better matching, “variance estimation must account for the fact that the same untreated subject may be in multiple matched sets.” [1]

Caliper matching is essentially the same as NN, but with the additional constraint that the pairs must be within a certain range (the caliper) of one another. This also ensures closer matching but can lead to reduced sample size, as unmatched participants must be excluded from the analysis. Radius matching is a variant of caliper matching which uses “not only the NN within each caliper but all of the comparison members within the caliper.” [4] The drawback with both methods, though, is that it’s difficult to know beforehand what a good caliper size will be. [7]

Results – example(s) of carrying out the methods in real situations like the group might encounter. A real or hypothetical data example and analysis can be effective in some cases. (2.5 points)

* Issues, Controversies, alternate approaches – any controversies or advantages/disadvantages to various approaches, alternate approaches. (1 point)

Summary, conclusions, and recommendations – take-away messages, key references. (1 point)

References

1. Austin, P.C., *An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies.* Multivariate Behav Res, 2011. **46**(3): p. 399-424.

2. Rosenbaum, P.R. and D.B. Rubin, *The central role of the propensity score in observational studies for causal effects.* Biometrika, 1983. **70**(1): p. 41-55.

3. Braitman, L.E. and P.R. Rosenbaum, *Rare outcomes, common treatments: Analytic strategies using propensity scores.* Annals of Internal Medicine, 2002. **137**(8): p. 693-695.

4. Caliendo, M. and S. Kopeinig, *Some Practical Guidance for the Implementation of Propensity Score Matching.* Journal of Economic Surveys, 2008. **22**(1): p. 31-72.

5. Lee, B.K., J. Lessler, and E.A. Stuart, *Improving propensity score weighting using machine learning.* Stat Med, 2010. **29**(3): p. 337-46.

6. Setoguchi, S., et al., *Evaluating uses of data mining techniques in propensity score estimation: a simulation study.* Pharmacoepidemiol Drug Saf, 2008. **17**(6): p. 546-55.

7. A. Smith, J. and P. E. Todd, *Does matching overcome LaLonde's critique of nonexperimental estimators?* Journal of Econometrics, 2005. **125**(1-2): p. 305-353.